Physica A 468 (2017) 53–69

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Exploring triad-rich substructures by graph-theoretic characterizations in complex networks

Songwei Jia [a,b], Lin Gao [a,*], Yong Gao [c], James Nastos [d], Xiao Wen [a], Xindong Zhang [a], Haiyang Wang [a]

[a] School of Computer Science and Technology, Xidian University, Xi'an 710071, China

[b] School of Software, Xidian University, Xi'an 710071, China

[c] Department of Computer Science, University of British Columbia, Okanagan, Kelowna, British Columbia, Canada V1V 1V7

[d] Department of Computer Science, Okanagan College, Kelowna, British Columbia, Canada BC V1Y 4X8

HIGHLIGHTS

- The novel triad-rich assumption.
- The definition of 2-club substructure.
- The edge niche centrality.
- The 2-hop overlapping strategy.
- The effective algorithm DIVANC.

ARTICLE INFO

Article history:
Received 8 March 2016
Received in revised form 27 July 2016
Available online 12 October 2016

Keywords:
Complex networks
Community detection
Triad-rich substructure
Graph-theoretic characterizations
Edge niche centrality
Overlapping

ABSTRACT

One of the most important problems in complex networks is how to detect communities accurately. The main challenge lies in the fact that traditional definition about communities does not always capture the intrinsic features of communities. Motivated by the observation that communities in PPI networks tend to consist of an abundance of interacting triad motifs, we define a 2-club substructure with diameter 2 possessing triad-rich property to describe a community. Based on the triad-rich substructure, we design a DIVision Algorithm using our proposed edge Niche Centrality DIVANC to detect communities effectively in complex networks. We also extend DIVANC to detect overlapping communities by proposing a simple 2-hop overlapping strategy. To verify the effectiveness of triad-rich substructures, we compare DIVANC with existing algorithms on PPI networks, LFR synthetic networks and football networks. The experimental results show that DIVANC outperforms most other algorithms significantly and, in particular, can detect sparse communities.

1. Introduction

One of the most important problems in complex networks is how to detect communities accurately [1]. Communities are the subsets of vertices with real physical sense. For example, in biological networks they are referred to as various biological functional modules such as protein complexes, GO terms and pathways; in social networks, communities may be various

* Corresponding author.
E-mail address: lgao@mail.xidian.edu.cn (L. Gao).

http://dx.doi.org/10.1016/j.physa.2016.10.021

social circles such as groups of people with common interests, etc. Traditional communities, which are typically described as dense subgraphs (subnetworks) explicitly or implicitly. The underlying assumption is that objects in some communities really tend to interact more frequently than in other regions of the network. Around the issue of how to detect communities, scholars have proposed many popular algorithms based on traditional community definitions which can identify parts of communities successfully at a certain degree. Examples of algorithms that detect communities by dense subnetworks include (i) random-walk based methods such as MCL [2] and INFOMAP [3]; (ii) seed-growing methods such as MCODE [4] and ClusterOne [5]; (iii) algorithms based on clustering, optimization, or statistical techniques such as LinkComm [6], LOUVAIN [7], and OSLOM [8]; and (iv) algorithms based on deeper graph-theoretic features such as EPCA [9–11].

While traditional definitions can offer some insight into some of the structure of communities, more and more recent studies show that these intuitions are unreliable [12–16]. Some of these examples include: overlapping communities have a higher density of links in the overlapping parts than in the non-overlapping ones, which are in contrast with the common picture of communities [16]; there is a paradox that the detection of well-defined communities is more difficult than the identification of ill-defined communities [12]. All of these counterintuitive evidences hint at the necessity of modifying the general defining characteristics of traditional communities. While there is a general consensus on the fact that there is a need for an adjustment of the notion of community or clusters, there is no clear direction to a remedy. Scholars [15] point out that there are two possible scenarios for filling the gaps between traditional definitions and communities. One is to include additional topological features in refining the traditional definitions beyond the standard measures of link density, degree correlations or density of loops, etc.; the other is to add requirements based non-topological knowledge, such as domain-specific background knowledge [17–19] for the detection of communities. However, in the former case, solely adjusting the structural conditions sought for may still not obtain satisfying results as the essence of communities in all contexts may not be characterized by equivalent topology. In the latter case, adding various domain-specific background knowledge may be effective on a limited number of cases, but the reliance on rigid domain-specific knowledge makes the resulting algorithms unlikely to exhibit scalability or transferability to other domains.

An ideal paradigm for fully analyzing communities would include identification of communities via certain specific intrinsic features, combined with a method for capturing deeper domain-specific structure in a general topological framework on which one can further develop algorithms. Here, we develop such a new framework by incorporating a novel and more subtle assumption based on graph-theoretic properties of communities and design efficient computing procedures to detect non-overlapping and overlapping substructures that have the desired properties. As shown in Fig. 1(a) and (b), both of the communities 'nuclear origin of replication recognition complex' [20] (dense) and 'GID complex' [21] (sparse) consist of abundantly interacting triad motifs, for instance. More details about the two complexes can be found in Appendix A. Motivated by the observation that communities in a PPI network are either quite dense or quite sparse and tend to consist of an abundance of interacting triad motifs [22–25], we define a 2-club substructure with diameter 2 possessing triad-rich property to describe a community. Based on the triad-rich substructure, we design a DIVision Algorithm using our proposed edge Niche Centrality DIVANC to detect communities effectively in complex networks. We also extend DIVANC to detect overlapping communities by proposing a simple 2-hop overlapping strategy. To verify the effectiveness of triad-rich substructures, we compare DIVANC with existing algorithms on PPI networks, LFR synthetic networks and football networks. The experimental results show that DIVANC outperforms most other algorithms significantly and, in particular, can detect sparse communities.

The rest of the paper is organized as follows. In Section 2, we present our framework for detecting communities. After discussing our datasets and providing statistical evidence that motivates and supports our triad-rich assumption about communities in Sections 2.1 and 2.2, we give a formal definition of a 2-club substructure in Section 2.3. In Section 2.4, we discuss the details of our algorithm for 2-club substructure detection, including a new edge-centrality measure specifically designed for 2-club substructures as well as a 2-hop-based strategy for extracting overlapping 2-club substructures. In Section 3, we report and discuss our experimental results. In Section 4, we conclude the paper and give closing discussion.

2. Methods

2.1. The datasets

We apply our framework on PPI networks [26,27], LFR synthetic networks [28,29] and football networks [30,31]. In the following, we give details about the relative networks and six golden standard sets of communities in PPI networks, respectively.

S. cerevisiae PPI networks (*Sce*DIP) are obtained from DIP [26] and *H. sapiens* PPI networks (*Hsa*HPRD) are extracted from HPRD [27]. For *Sce*DIP, we use the sets from the Munich Information Center for Protein Sequences (MIPS) [32], Saccharomyces Genome Database (SGD) [33] and *S. cerevisiae* GO terms (*Sce* GO term) as golden standards [34,35]. For *Hsa*HPRD, we use the sets of Human Protein Complex Database with a Complex Quality Index (PCDq) [36], Comprehensive Resource of Mammalian Protein Complexes (CORUM) [37] and *H. sapiens* GO terms (*Hsa* GO term) [34,35] as golden standards. *Sce*DIP consists of 4980 proteins and 22 076 interactions; *Hsa*HPRD consists of 9269 proteins and 36 917 interactions. The GO terms are not composed of all the terms but the high-level GO terms whose information content is more than 2 [34,35]. The definition of the information content (IC) of a GO term g is $IC = -\log(|g| / |root|)$ as given in the literature [34], where '*root*' is the corresponding root GO terms across the three aspects of molecular function (MF), biological process (BP) or cellular

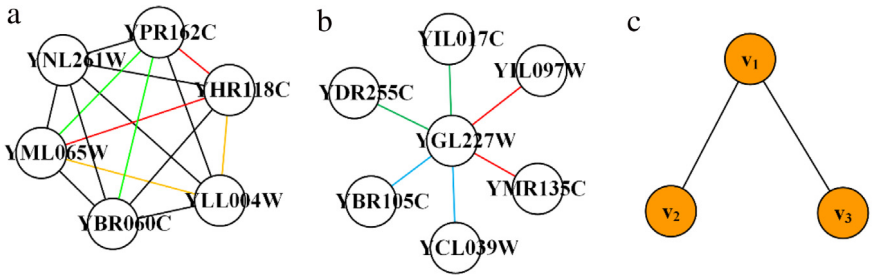


Fig. 1. Communities consisting of abundantly interacted triad motifs. (a) Nuclear origin of replication recognition complex; (b) GID complex; (c) an example of triad motif.

component (CC) of g. In addition, the GO terms with less than 2 proteins are removed. We also remove the protein complexes or GO terms of which no members appear in the corresponding PPI networks. Last, MIPS consists of 203 and SGD has 305 protein complexes, while PCDq includes 1204 and CORUM has 1294 complexes. Additionally, there are 1050 terms in *Sce* GO term, and 4457 terms in *Hsa* GO term.

Here we also give the details of LFR synthetic networks [28,29]. The parameters of the series of LFR networks are: vertices size $N = 1000$, average degree $\bar{k} = 15$, minimum community size $minc = 20$, maximum community size $maxc = 50$, the mixing parameter *mu* with a step of 0.1 from 0.1 to 1.0, and for overlapping LFR networks, additional parameters such as number of overlapping vertices $on = 100$, number of memberships of the overlapping vertices $om = 2$. The parameters were chosen to follow the examples provided by the original code and we downloaded it at http://santo.fortunato.googlepages.com/inthepress2.

Football network [30,31] represents the relationships played among college teams during the year 2001 football season of the USA, and consists of 115 vertices and 613 edges, indicating 115 teams and 613 games played against each other. The 115 teams are grouped into 11 conferences, with a 12th group of independent teams.

2.2. Triad-rich substructures: a novel assumption on communities

As mentioned in Section 1, most existing community detection algorithms are (explicitly or implicitly) based on the assumption that communities most likely appear in dense subnetworks. This edge-rich assumption results in two fundamental difficulties that make it hard, if not impossible, to improve the performance of those methods that detect communities by extracting dense subnetworks: (1) the requirement for a subnetwork to be highly dense is too strong; and (2) a pure density-based measure cannot distinguish among subnetworks that have different internal structures that may be of physical significance.

To further evidence the rough assumption that communities most likely appear in dense subnetworks is not comprehensive enough, we quantitatively analyze the density distributions of the communities among six golden standard sets in corresponding PPI networks, for instance. The numbers of elements in golden standard sets of PPI networks are described in Table 1. As shown in Fig. 2, we demonstrate the percentages of communities among their whole golden standard sets of SGD, MIPS, *Sce* GO term, PCDq, CORUM and *Hsa* GO term according to their density distributions, respectively. We consider the percentages of communities with their densities 0, greater than 0 but no more than 0.1, greater than 0.1 but no more than 0.2,…, densities 1 but sizes greater than 2, densities 1 but sizes 2 respectively, as described in the legend of Fig. 2. Here, we demonstrate the percentages of those with densities 1 but sizes 2 since although these communities seem very dense, they are merely paths with two vertices. As shown in Fig. 2, there are only a small number of communities with high density in the golden standard sets of SGD, MIPS, *Sce* GO term, PCDq, CORUM and *Hsa* GO term.

Motivated by the observation that communities (either dense or sparse) in a PPI tend to consist of abundantly interacting triad motifs [22–25], we propose that a community detection method shall be based on the following triad-rich assumption: communities are most likely to occur in the substructures that contain many interacting triads. A motif in complex networks is a pattern of subnetworks on a small number of vertices that occur at a significantly higher frequency than what is expected in a random network with similar network statistics. In this paper, we will focus on the most basic blocks of triad motif [38] that consist of 3-vertices and two links, as depicted in Fig. 1(c). This is because (i) a 2-vertices motif is nothing but an edge, and is trivial; and (ii) motifs containing more than three vertices can be constructed from interacting triad motifs.

The triad-rich assumption naturally generalizes the edge-rich assumption in that a dense subnetwork is triad-rich. For example, the 'Nuclear origin of replication recognition complex' shown in Fig. 1(a) is a clique with the largest density, it is not hard to see that the complex contains many interacting triad motifs. However, a triad-rich subnetwork is not necessarily edge-rich, making it possible to detect communities that are not necessarily dense. For example, the 'GID complex' shown in Fig. 1(b) is a star subnetwork with the lowest density but containing many interacting triad motifs. To quantify the property of being triad-rich, we impose the requirement that every pair of vertices in a community participates in at least one triadic interaction. This leads to the graph-theoretic definition of a triad-rich community as a substructure with diameter 2 and triad-rich property (i.e., a 2-club). This definition of a triad-rich substructure makes it possible to study interesting internal structures of communities, which cannot be distinguished by any density-mainly measure.

Table 1
The details of six golden standard sets and their corresponding *P*-values of triad distribution in PPI networks.

Golden standards	MIPS	SGD	PCDq	CORUM	*Sce* GO term	*Hsa* GO term
Numbers	203	305	1204	1294	1050	4457
P-value	6.98e−08	3.69e−05	2.49e−15	2.46e−18	4.79e−11	1.97e−52

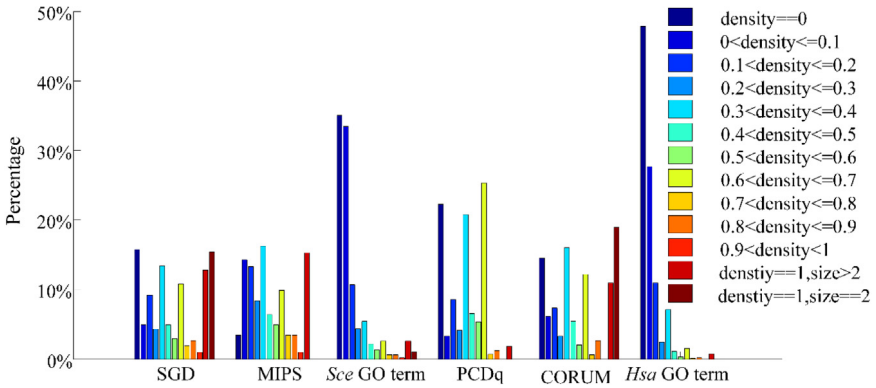


Fig. 2. Density distribution among six golden standard sets in PPI networks.

To further support the proposed triad-rich assumption that communities are more likely to occur in substructures that contain many interacting triads, we give a basic statistical analysis on the distribution of triad motifs in communities on PPI networks. For PPI networks, we use the above six golden standard sets described in Section 2.1 as communities, we compare the frequencies of triad motifs in communities to those of a random selection of equally-sized subnetworks. Our process is as follows: we first count the number of triads existing in communities. For each of the communities, if it contains n vertices, we randomly choose a set of n vertices in the corresponding PPI network and count the triad motifs among those randomly-selected vertices. We repeat this random selection one thousand times and calculate the corresponding average triad number. Thus for each of the golden standards, we obtain a pair of vectors, one of which indicates the triad numbers for the communities and the other indicating the average triad numbers for the subnetworks obtained by randomly choosing vertices. The dimensions of the vectors are equal to the numbers of communities in their golden standard sets. For instance, for the MIPS, we have a vector of 203 values of the true counts of triad motifs in the 203 complexes, along with a vector of triad counts in randomly generated subnetworks of equal size to each of the complexes. To test the statistical significance for the triad distributions among each of the golden standard sets, we calculate the corresponding P-values based on a T-test by comparing the number of triads obtained in the communities to the numbers obtained in the randomly-selected equal-sized subnetworks. The lower P-values mean the more significant triad distribution in communities. We display the corresponding P-values for MIPS, SGD, PCDq, CORUM, *Sce* GO term and *Hsa* GO term in Table 1 respectively, where it is readily seen that the randomly-selected subnetworks have statistically fewer triad motifs than the true benchmarks. Thus, triad motifs are distributed far more densely in communities than in randomly-selected subnetworks. This result reinforces our proposed novel assumption that communities consist of abundantly interacted triad motifs.

2.3. A graph-theoretic definition of triad-rich substructures

In this section, we mainly introduce relative terminologies and our graph-theoretic definition of a triad-rich substructure.

2.3.1. Terminologies and concepts in graph theory

A graph or network $G = (V, E)$ consists of a vertex set V and an edge set E. An induced subgraph of a graph is specified by a set of vertices, and all of the edges that exist on those vertices in the network are also part of the induced subgraph [39]. A P_4 is an induced graph on four ordered vertices, which are connected as a simple path [11,39]. The distance between two vertices is the length (i.e., the number of edges) of a shortest path between them. The diameter of a graph is the maximum distance between a pair of vertices.

2.3.2. 2-club substructures

We define a triad-rich substructure to be an induced and connected substructure where every pair of vertices participates in at least one triadic interaction. It is not hard to see that a substructure is a triad-rich one if and only if it is an induced subgraph of diameter 2. Noting that a diameter-2 induced subgraph is also known as a 2-club in the social network literature, we shall call our triad-rich substructure a 2-club substructure. In a 2-club substructure, every pair of vertices either forms

an edge or is contained in at least one triad motif and in fact, 2-club substructures play the same role in the class of triad-rich subnetworks as cliques do in the class of edge-rich subnetworks.

2.4. DIVANC: a division algorithm for finding 2-club substructures

Given the definition of a triad-rich substructure as a 2-club, a natural algorithmic problem is to delete the minimum number of edges to modify a network into a new network where each connected component has diameter 2. While we have not been able to give a formal proof, we believe that this problem is NP-hard, similar to the many NP-hard edge-deletion problems. In this section, we develop a new centrality measure to approximate the requirement of being of diameter 2, and design effective and efficient algorithm for detecting both non-overlapping and overlapping 2-club substructures. Our algorithm is an edge division algorithm that removes edges according to a new edge centrality measure, called the edge niche centrality, specifically designed to capture the properties of 2-club substructures.

2.4.1. Edge niche centrality

In their seminal work, Girvan and Newman [30] proposed an edge-division algorithm to detect communities by iteratively removing edges with high edge betweenness centrality. One of the issues with the G–N algorithm is the high time complexity of computing the edge betweenness centrality, even though there are polynomial time algorithms for it. Recently, a few easy-to-compute centrality measures have been proposed and used to design more efficient edge-division algorithms, including the P_4 centrality [11], anti-triangle centrality [40], and the edge clustering coefficient [41].

The new edge-centrality measure, edge niche centrality, measures the importance of an edge by taking into consideration the edge's P_4 centrality and embeddedness (revealed by edge clustering coefficient). In the following, we give the formal definition of our edge niche centrality. Let $G = (V, E)$ be an undirected and unweighted network and e_{ij} be an arbitrary edge in G, the niche centrality C_{ij}^N of e_{ij} is defined as:

$$C_{ij}^N = C_{ij}^- + \min\left(k_i - 1, k_j - 1\right) / \left(C_{ij}^\Delta + 1\right) \tag{1}$$

where C_{ij}^- represents the edge P_4 centrality defined as the number of P_4s which e_{ij} belongs to, and can be calculated by the function *Is* P_4 (a, b, c, d) provided in Ref. [11] simply; C_{ij}^Δ is the number of triangles which e_{ij} belongs to, representing the embeddedness of e_{ij} (i.e., the number of common neighbors of vertices v_i and v_j); k_i (k_j) denotes the degree of the vertex v_i (v_j).

As shown in Eq. (1), two factors are considered in the edge niche centrality. The P_4 centrality, helps in identifying edges that participate in many induced paths of length 3. Removing edges with high P_4 centrality helps separating vertices that have distance greater than 2 and therefore, are not likely to be in the same 2-club substructure. The second term distinguishes edges that have similar P_4 centrality, but have different embeddedness. This definition of edge niche centrality gives us a way to quantitatively measure the extent to which an edge is an inter-link or intra-link. If its niche centrality is large the edge is more likely to be an inter-link, while if its niche centrality is smaller it is more likely to be an intra-link.

2.4.2. The 2-hop overlapping strategy

As edge-division algorithms can only detect non-overlapping substructures, we propose a strategy, the 2-hop overlapping strategy, to uncover 2-club substructures that may overlap. Our strategy is inspired by the idea of overlapping communities in Ref. [42]. It searches eligible peripheral vertices and adds them into non-overlapping substructures to obtain the corresponding overlapping 2-club substructures. The criterion used to add a peripheral vertex is based on its closeness to a 2-club substructure. Formally, for a given non-overlapping 2-club substructure $M = (V_M, E_M)$ from $G = (V, E)$, the set of vertices to be added into V_M (denoted as $AVS\ (V_M)$) is defined to be

$$AVS\ (V_M) = \left\{ v_v \left| \forall v_x \in V_M,\ gd\ (v_x, v_v) \le 2,\ \text{and}\ \left| N(v_v) \bigcap V_M \right| / |V_M| > 0.5, v_v \in N\ (M) \right. \right\} \tag{2}$$

where $gd\ (v_x, v_v)$ representing the distance between vertices v_x and v_v, the neighborhood sets of M and v_v are $N\ (M) = \{v_u\ |(v_u, v_v) \in E, v_v \in V_M, v_u \in V, v_u \notin V_M\}$, and $N(v_v) = \{v_u\ |(v_u, v_v) \in E, v_u \in V\}$.

Note that the new subnetwork on the vertex set $V_M \bigcap AVS(V_M)$ is still a 2-club substructure. This is because that any vertex in $AVS(V_M)$ must be of distance at most 2 to every vertex in V_M and that every pair of vertices in $AVS(V_M)$ has a common neighbor in V_M. An example as shown in Fig. 3, the vertex v_f is the overlapping vertex belonging to the two substructures M_1 and M_2 simultaneously since we have $v_f \in AVS\left(V_{M_1}\right)$ and $v_f \in AVS\left(V_{M_2}\right)$, where the vertices $v_g \notin AVS(M_1)$ since $\left|N(v_g) \bigcap V_{M_1}\right| / \left|V_{M_1}\right| < 0.5$, $v_h \notin AVS(M_1)$ since $gd\ (v_e, v_h) > 2$ and $v_m \notin AVS(M_1)$ since even $v_m \notin N(M_1) = \{v_f, v_g, v_h\}$.

2.4.3. Details of DIVANC

As shown in Table 2, DIVANC removes edges iteratively according to their edge niche centrality until all the connected components are 2-club substructures. If needed, DIVANC can include an additional step to construct overlapping 2-club substructures by using the 2-hop overlapping strategy.

The effectiveness in practice of DIVANC will be reported in Section 3. In the following, we discuss its worst-case complexity. Let $\bar{k}$ be the average degree of the vertices and T the number of edges removed. If the overlapping step is not

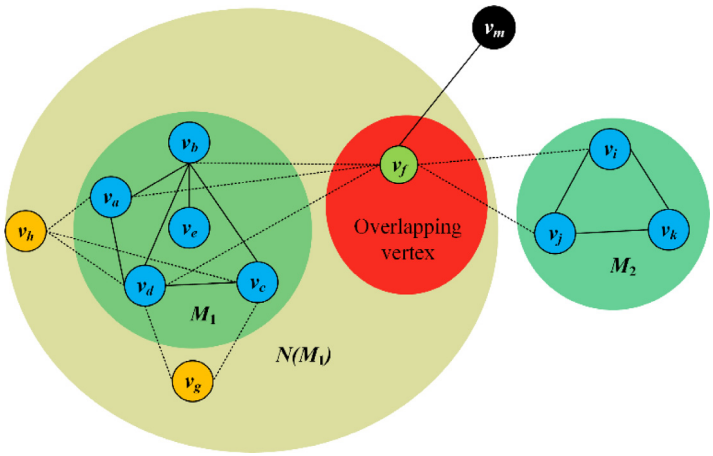


Fig. 3. An example for demonstrating the 2-hop overlapping strategy. Vertex v_f is the overlapping vertex searched by the 2-hop overlapping strategy, which belongs to the substructures M_1 and M_2.

Table 2
The diagram of DIVANC.

Algorithm DIVANC
Input: network $G = (V, E)$; **Output**: 2-club substructures; 1: Calculate the edge niche centrality score for each edge of $G = (V, E)$; 2: ***While*** there are connected components of diameter greater than 2 ***do*** 3: Remove the edge with the highest niche centrality score among all the edges; 4: Re-calculate the scores of those edges affected by the removal of the edge; 5: ***End while*** 6: ***If*** overlapping 2-club substructures needed ***then*** 7: ***For*** each of the current connected components ***do*** 8: Apply the 2-hop overlapping strategy; 9: ***End for*** 10: ***End if***

performed, the time complexity of DIVANC is $O(\bar{k}^2\,|E| + \bar{k}^4 T)$ where the first term is the time to compute the edge niche centrality for all edges and the second term is the time to remove the T edges. When the overlapping step is performed, the total running time is $O(\bar{k}^2\,|E| + \bar{k}^4 T + \bar{k}\,|V|)$. Since general practical networks usually have a small average degree $\bar{k}$ and since T is at most the sum of $|E| - (|V| - 1)$ and the number of detected 2-club substructures, the running time of DIVANC is very low, thus it is very efficiently. We note that neither T nor the number of detected 2-club substructures is a parameter of the algorithm.

3. Experiments and analyses

We report our experiment results and their analyses in this section. In Section 3.1 we compare DIVANC with existing algorithms in the literature on the PPI networks, LFR synthetic networks and football networks to verify the effectiveness of triad-rich substructures. We show the advantage of DIVANC in detecting sparse communities in Section 3.2. In Section 3.3 we test the practical performance of our proposed edge niche centrality and 2-hop overlapping strategy.

3.1. Verifying the effectiveness of triad-rich substructures

In this section we mainly compare DIVANC with other widely-used reference algorithms to test the effectiveness of triad-rich substructures. To compare fairly, we select corresponding competing non-overlapping and overlapping algorithms respectively since DIVANC can also be extended into overlapping version, which is denoted as DIVANC′ temporarily for comparison. Other than the well-known community detection algorithms, we also choose some excellent domain-specific algorithms (detecting protein complexes) such as COACH [43] and ClusterOne [5].

Among the non-overlapping algorithms, we freely downloaded the Cytoscape plugin for MCODE [4] at http://www.cytoscape.org/. We implemented INFOMAP [3] freely by the *R* package igraph [44]. We obtained the source code for MCL at http://www.micans.org/mcl. We obtained the code of MATLAB version of LOUVAIN [7] at http://perso.uclouvain.be/vincent.blondel/research/louvain.html. EPCA [11] for detecting the defined cograph communities based on the edge P_4 centrality, which is one critical component of the edge niche centrality proposed in this paper and we have the code. Especially, in

order to verify the delicate advantages of edge niche centrality over that of edge P_4 centrality, we scrabble up a special edge division algorithm based on edge P_4 centrality to detect 2-club substructures (temporarily denoted as EPD2, Edge P4 centrality and Diameter 2 stop criterion) like our DIVANC. Thus, the different effectiveness of DIVANC and EPD2 can be just due to their own different edge centralities.

While, as for the overlapping algorithms, we freely downloaded the Cytoscape plugin for ClusterOne [5] at http://www.cytoscape.org/. We obtained the executable program for COACH [43] at http://www1.i2r.a-star.edu.sg/~xlli/. We used LinkComm [6] by its *R* package [45]. We made use of its fast version OSLOM2 [8] at http://oslom.org/software.htm. We set all the corresponding parameters of those competing algorithms at their respective default values as they report that the algorithms can obtain best performances under default parameter values. Especially inspired by scrabbling up the special algorithm EPD2 among the non-overlapping algorithms, in this section we further extend EPD2 into its overlapping version based the proposed 2-hop overlapping strategy, which is denoted as EPD2′. Introducing EPD2′ as a competing algorithm can not only provide further comparative perspective between edge niche centrality and P_4 centrality in an overlapping context, but also can verify the portability of the 2-hop overlapping strategy.

The effectiveness of those algorithms is evaluated using a series of indices in terms of protein complex detection and GO term detection. We use the indices of the numbers of matching communities, the cluster-wise sensitivity (*Sn*), cluster-wise positive predictive value (*PPV*), the accuracy score (*Acc*), maximum matching ratio (MMR) [5,43,46] to assess the algorithms in complexes detection. *F*-measure and percentage of matched GO terms and MMR are used to assess them in identifying GO terms [34,35]. More details about the used indices can be found in Appendix B.

3.1.1. Comparison in detecting protein complexes and GO terms on PPI networks

The results on the effectiveness for detecting protein complexes of non-overlapping algorithms are summarized in Table 3 and overlapping ones in Table 4. Among the indices, we mainly pay more attention to the three indices: numbers of candidate complexes which can match at least one reference complex among golden standards (NMC), the accuracy scores (*Acc*) and maximum matching ratio (MMR), as given in bold fonts in Tables 3 *and* 4. In addition to comparing them in detecting protein complexes, we also compare their effectiveness in detecting GO terms. We test the compared algorithms for detecting GO terms from *Sce*DIP and *Hsa*HPRD using the indices of *F*-measure, percentage of matched GO terms and maximum matching ratio. Fig. 4((a)–(c)) show the indices of *F*-measure, percentage of matched GO terms and maximum matching ratio for non-overlapping algorithms on *Sce*DIP and *Hsa*HPRD respectively. Fig. 4((d)–(f)) display the corresponding indices for overlapping algorithms.

As shown in Table 3, among the non-overlapping algorithms, DIVANC has the largest numbers of matched protein complexes across all the golden standards except PCDq. Where it has 377 matched protein complexes, which is almost equal to the highest number 378. The maximum matching ratios of DIVANC are the highest ones on SGD and CORUM, and while across MIPS and PCDq the maximum matching ratios of DIVANC are very close to the highest ones. The accuracy scores of DIVANC are also very close to their highest ones such as those of MCL, EPCA and EPD2. As demonstrated in Table 4, among the overlapping algorithms, DIVANC′ has the largest accuracy scores across all the golden standards. Except on PCDq DIVANC has the largest maximum matching ratio, on other golden standards the maximum matching ratios of DIVANC are lower than those of COACH. As Fig. 4 shows, the bar plots for illustrating the effectiveness in GO terms detection also clearly reveal that DIVANC, MCL are competitive among non-overlapping algorithms, while among overlapping algorithms DIVANC′, COACH, LinkComm are competitive and they all outperform others like the instances about complexes detection described in Tables 3 and 4. The reason for COACH and LinkComm possessing better effectiveness than DIVANC′ in detecting GO terms is that both of COACH and LinkComm can obtain highly overlapping communities, while DIVANC′ can just obtain periphery overlapping 2-club substructures. Thus in the further research on the one hand we should continue maintaining the unique graph-theoretic characteristics of 2-club substructures, on the other hand we should pay more attention to improving their overlapping extent.

To give a specific example, we especially select a simple complex named CCBL2-HBXIP-RABIF-UTP14A complex, which can be detected perfectly by DIVANC′ but cannot by other algorithms. As shown in Fig. 5, the CCBL2-HBXIP-RABIF-UTP14A complex consists of four-subunit proteins, which is a protein complex stored in an integrated database of human genes and transcripts, the H-Invitational Database (H-InvD) [47]. The proteins in green color are the members of CCBL2-HBXIP-RABIF-UTP14A complex and those in dark red color are not. We emphasize that among all the non-overlapping and overlapping algorithms, only DIVANC′ can detect the CCBL2-HBXIP-RABIF-UTP14A complex perfectly, while none of the algorithms MCODE, INFOMAP, LinkComm and OSLOM2 can detect meaningful candidate complex successfully, not to mention matching perfectly with the benchmark, more details in table S1.

3.1.2. Comparing the algorithms on LFR synthetic networks

Although as the foundation of our 2-club substructures framework, the triad-rich substructures assumption about communities are observed from PPI networks, what we want to emphasize is that either DIVANC or DIVANC′ can work well on general complex networks. In the following we mainly test the scalability of DIVANC on LFR synthetic networks [28,29]. In the testing experiments, we use the well-known normalized mutual information (NMI) [48,49] (more details see in Appendix B) for evaluating community detection algorithms.

Table 3
Comparison with non-overlapping algorithms for detecting complexes from *Sce*DIP and *Hsa*HPRD.

Net[a]	GS[b]	Alg[c]	Cov[d]	NM[e]	AS[f]	NMC[g]	*Sn*	*PPV*	*Acc*	MMR
SceDIP	MIPS		1061	203	12.52	–	–	–	–	–
		MCODE	781	51	15.31	15	0.2149	0.1987	0.2066	0.0301
		INFOMAP	4980	441	11.29	47	0.4915	0.3190	**0.3960**	0.0961
		MCL	4736	928	5.10	69	0.3125	0.3689	0.3395	0.1666
		LOUVAIN	4980	675	7.38	35	0.5081	0.2571	0.3614	0.0849
		EPCA	4687	1019	4.60	82	0.3530	0.3982	0.3749	**0.2006**
		EPD2	4723	1015	4.65	82	0.3589	0.3984	0.3782	0.1980
		DIVANC	4856	1128	4.30	**88**	0.3526	0.4008	0.3759	0.2004
	SGD		1211	305	5.70	–	–	–	–	–
		MCODE	781	51	15.31	21	0.3076	0.2490	0.2768	0.0358
		INFOMAP	4980	441	11.29	74	0.6354	0.4447	0.5316	0.1050
		MCL	4736	928	5.10	124	0.5026	0.5585	0.5298	0.1884
		LOUVAIN	4980	675	7.38	79	0.6538	0.3598	0.4850	0.1259
		EPCA	4687	1019	4.60	129	0.5348	0.5943	**0.5638**	0.2073
		EPD2	4723	1015	4.65	128	0.5382	0.5924	0.5647	0.2023
		DIVANC	4856	1128	4.30	**141**	0.5348	0.5918	0.5626	**0.2108**
*Hsa*HPRD *PCDq*			3433	1204	4.51	–	–	–	–	–
		MCODE	1121	100	11.21	27	0.1624	0.1816	0.1717	0.0990
		INFOMAP	9269	668	13.88	150	0.5192	0.3266	0.4118	0.0480
		MCL	8903	1789	4.98	316	0.3992	0.5322	0.4609	0.1246
		LOUVAIN	9269	1097	8.45	226	0.5385	0.2944	0.3981	0.0962
		EPCA	8807	1946	4.53	377	0.3856	0.5504	0.4607	**0.1450**
		EPD2	8855	1942	4.56	**378**	0.3872	0.5491	**0.4611**	0.1448
		DIVANC	9077	2151	4.22	377	0.3804	0.5587	0.4610	0.1443
	CORUM		1955	1294	5.06	–	–	–	–	–
		MCODE	1121	100	11.21	23	0.2452	0.0791	0.1392	0.0087
		INFOMAP	9269	668	13.88	73	0.5251	0.1591	0.2890	0.0210
		MCL	8903	1789	4.98	190	0.4041	0.2460	**0.3153**	0.0613
		LOUVAIN	9269	1097	8.45	95	0.5663	0.1310	0.2724	0.0327
		EPCA	8807	1946	4.53	196	0.3772	0.2529	0.3088	0.0642
		EPD2	8855	1942	4.56	196	0.3810	0.2528	0.3103	0.0639
		DIVANC	9077	2151	4.22	**231**	0.3735	0.2599	0.3116	**0.0723**

[a] Net Networks.
[b] GS Golden standards.
[c] Alg Algorithms.
[d] Cov Numbers of coverage proteins.
[e] NM Numbers of detected candidate complexes.
[f] AS Average size of obtained candidate complexes.
[g] NMC Numbers of candidate complexes which can match at least one reference complex.

The testing LFR synthetic networks include a series of non-overlapping networks and overlapping networks respectively. The parameters for producing LFR non-overlapping and overlapping synthetic networks are introduced in Section 2.1. We compare the NMI values of the results obtained by the compared non-overlapping and overlapping algorithms on the synthetic networks as shown in Fig. 6. Each node of the figure corresponds to the average NMI value over 20 LFR networks produced on the same parameters. The NMI values of all algorithms decrease as the mixing parameter *mu* increases. The reason is that community structures of the LFR networks become fuzzier and fuzzier, and thus are more difficult to be detected correctly as *mu* increases. As Fig. 6(a) shows, the purple line with diamond signs represents the NMI value of DIVANC and Fig. 6(b) shows that the purple line with cross signs represents that of DIVANC′. Moreover, the results of the rest other algorithms are indicated by the corresponding color lines with corresponding signs as shown in Fig. 6. INFOMAP can obtain the best effectiveness among these compared non-overlapping algorithms and OSLOM2 has the highest NMI value among those overlapping algorithms. As Fig. 6 shows, DIVANC has competitive effectiveness among the non-overlapping algorithms and the second highest NMI value among overlapping algorithms. DIVANC obviously outperforms EPCA [11] and MCODE [4], while DIVANC′ has better effectiveness than LinkComm [6], ClusterOne [5], and COACH [43]. Both of DIVANC and DIVANC′ can obtain competitive effectiveness on LFR synthetic networks reveal to us that the proposed 2-club substructure is suitable for synthetic networks at certain extent, but we really need to improve it since the triad-rich assumption is observed just only from PPI networks.

3.1.3. Comparison on football networks

In this section we also test them on a small social network the widely-used football networks [30,31]. As introduced in Section 2.1, football network consists of 115 teams and 613 games and the 115 teams are grouped into 11 conferences, with a 12th group of independent teams (without obvious affiliations, we artificially arrange the 8 independent teams into the 12th group together for convenience) as shown in Fig. 7(a). We display their NMI values of the compared non-overlapping and overlapping algorithms respectively in Table 5. DIVANC and DIVANC′ can obtain the same result. The NMI value of DIVANC is

Table 4
Comparison with overlapping algorithms for detecting complexes from *Sce*DIP and *Hsa*HPRD.

Net[a]	GS[b]	Alg[c]	Cov[d]	NM[e]	AS[f]	NMC[g]	*Sn*	*PPV*	*Acc*	MMR
SceDIP	MIPS		1 061	203	12.52	–	–	–	–	–
		COACH	7 814	886	8.82	**165**	0.3790	0.2781	0.3246	**0.2831**
		ClusterOne	2 218	596	3.72	76	0.2645	0.3815	0.3176	0.1424
		LinkComm	6 587	875	7.53	156	0.4176	0.3299	0.3711	0.2299
		OSLOM2	5 442	85	64.02	21	0.5053	0.2382	0.3469	0.0310
		EPD2′	4 986	1015	4.91	89	0.3754	0.3845	0.3800	0.2121
		DIVANC′	5 129	1128	4.55	96	0.3896	0.3862	**0.3879**	0.2221
	SGD		1 211	305	5.70	–	–	–	–	–
		COACH	7 814	886	8.82	**232**	0.5509	0.3774	0.4560	**0.2731**
		ClusterOne	2 218	596	3.72	126	0.4158	0.5941	0.4970	0.1767
		LinkComm	6 587	875	7.53	213	0.5543	0.3989	0.4703	0.2123
		OSLOM2	5 442	85	64.02	24	0.6475	0.2745	0.4216	0.0298
		EPD2′	4 986	1015	4.91	133	0.5566	0.5629	0.5597	0.2107
		DIVANC′	5 129	1128	4.55	150	0.5716	0.5626	**0.5671**	0.2272
*Hsa*HPRD *PCDq*			3 433	1204	4.51	–	–	–	–	–
		COACH	14 086	725	8.17	**422**	0.3937	0.1584	0.2497	0.1096
		ClusterOne	4 151	1103	3.76	286	0.2591	0.6746	0.4181	0.1029
		LinkComm	11 194	1605	6.97	330	0.3750	0.2958	0.3331	0.0826
		OSLOM2	10 016	208	48.15	19	0.5262	0.1686	0.2978	0.0068
		EPD2′	9 300	1942	4.79	402	0.4088	0.4506	0.4292	**0.1507**
		DIVANC′	9 626	2151	4.48	401	0.4058	0.4564	**0.4304**	**0.1507**
	CORUM		1 955	1294	5.06	–	–	–	–	–
		COACH	14 086	1725	8.17	**443**	0.4653	0.0681	0.1780	**0.1103**
		ClusterOne	4 151	1103	3.76	164	0.2711	0.2780	0.2745	0.0548
		LinkComm	11 194	1605	6.97	372	0.4308	0.1342	0.2404	0.0910
		OSLOM2	10 016	208	48.15	20	0.5425	0.0970	0.2294	0.0057
		EPD2′	9 300	1942	4.79	213	0.4201	0.2113	0.2980	0.0697
		DIVANC′	9 626	2151	4.48	254	0.4233	0.2131	**0.3004**	0.0800

The foot note of Table 4 being the same as that of Table 3.

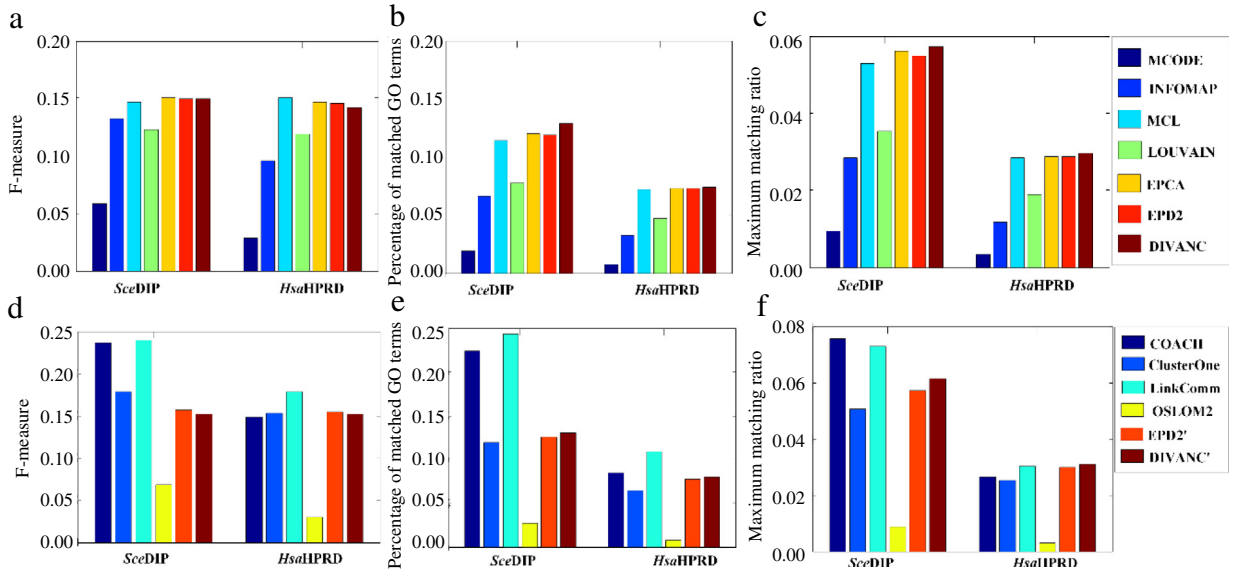
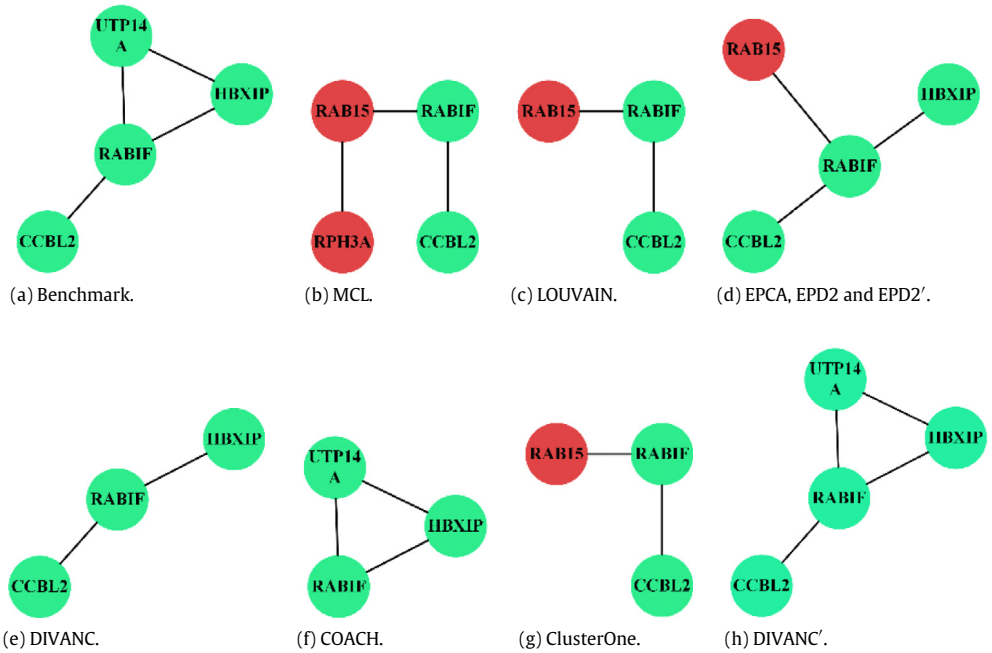


Fig. 4. The bar plots illustrating the effectiveness of non-overlapping and overlapping algorithms for detecting GO terms. Fig. 4((a)–(c)) demonstrating the indices of *F*-measure, percentage of matched GO terms and maximum matching ratio of non-overlapping algorithms on *Sce*DIP and *Hsa*HPRD respectively; Fig. 4((d)–(f)) displaying the corresponding indices of overlapping algorithms.

in close proximity to the highest one of LOUVAIN among non-overlapping algorithms, while among overlapping algorithms DIVANC′ gains the highest NMI value. DIVANC gains 12 2-club substructures after removing 190 edges. Surprisingly, we find the 12 2-club substructures matching the 12 real football conferences in a nearly perfect way as shown in Fig. 7(b). Other than three of the 8 independent teams presented by green triangles as shown in Fig. 7(a) are misarranged just since they are the independent teams without obvious affiliations, all of the rest teams match the real groups perfectly. As shown in Fig. 7(b), two independent teams Navy and Notre Dame are arranged into the green circle group and another independent



(a) Benchmark. (b) MCL. (c) LOUVAIN. (d) EPCA, EPD2 and EPD2′.

(e) DIVANC. (f) COACH. (g) ClusterOne. (h) DIVANC′.

Fig. 5. Illustration of the candidate complexes detected by the competing algorithms about CCBL2-HBXIP-RABIF-UTP14A complex. Fig. 5(a) the benchmark of CCBL2-HBXIP-RABIF-UTP14A complex and Fig. 5((b)-(h)) the corresponding candidate complexes detected by the non-overlapping algorithms MCL, LOUVAIN, EPCA, EPD2, DIVANC, and the overlapping algorithms COACH, ClusterOne, EPD2′, DIVANC′, where the genes in green color being the members of CCBL2-HBXIP-RABIF-UTP14A complex and those in dark red color not. Unfortunately, none of the algorithms MCODE, INFOMAP, LinkComm and OSLOM2 are being able to detect valuable candidate complex successfully. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

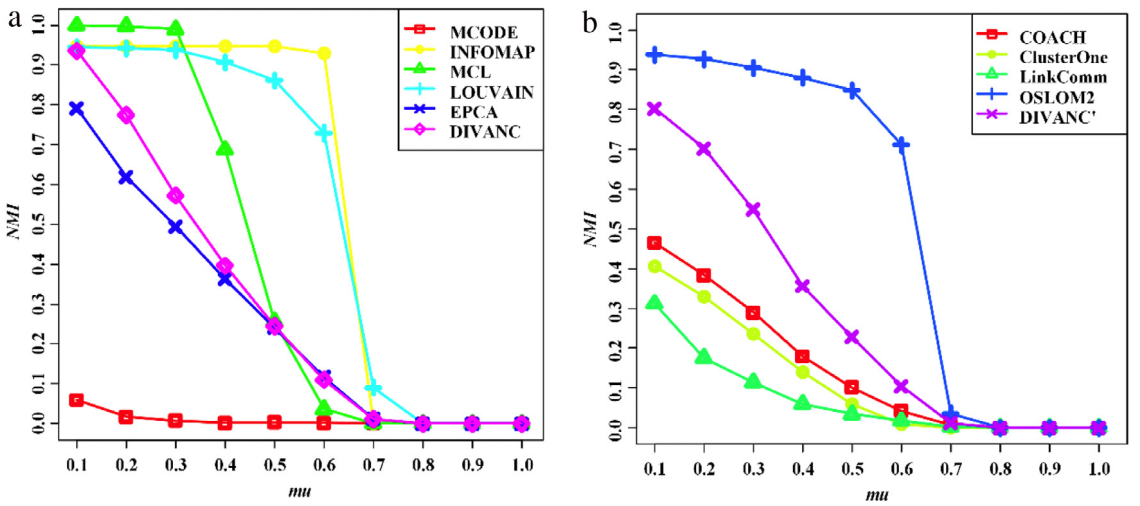


Fig. 6. Illustration of the average NMI values of the results obtained by the compared algorithms on the series of LFR networks as *mu* from 0.1 to 1.0 with a step of 0.1. Fig. 6(a) the average NMI values of non-overlapping algorithms; Fig. 6(b) the average NMI values of overlapping algorithms.

teams Connecticut is partitioned into the red triangle group irrelevantly. Notably DIVANC has signally better effectiveness than EPCA [11] since there are no isolated vertices among the obtained 2-club substructures, deleting 190 edges much lower than 290 ones that of EPCA and just only three misarranged teams, much fewer than that of EPCA. It is obvious that our algorithm also has impressive effectiveness on football networks.

3.2. The advantage for detecting sparse communities

Other than the above macroscopic comparisons according various indices, in this section due to the triad-rich assumption that underpins our definition of 2-club substructures, we show the advantage of DIVANC for detecting sparse communities.

Table 5
The NMI for effectiveness comparison on football networks.

Non-overlapping algorithms	NMI of non-overlapping algorithms	Overlapping algorithms	NMI of overlapping algorithms
MCODE	0.3834	COACH	0.5861
INFOMAP	0.8332	ClusterOne	0.6064
MCL	0.8332	LinkComm	0.2814
LOUVAIN	**0.8361**	OSLOM2	0.8150
EPCA	0.6917	DIVANC′	**0.8332**
DIVANC	0.8332	–	–

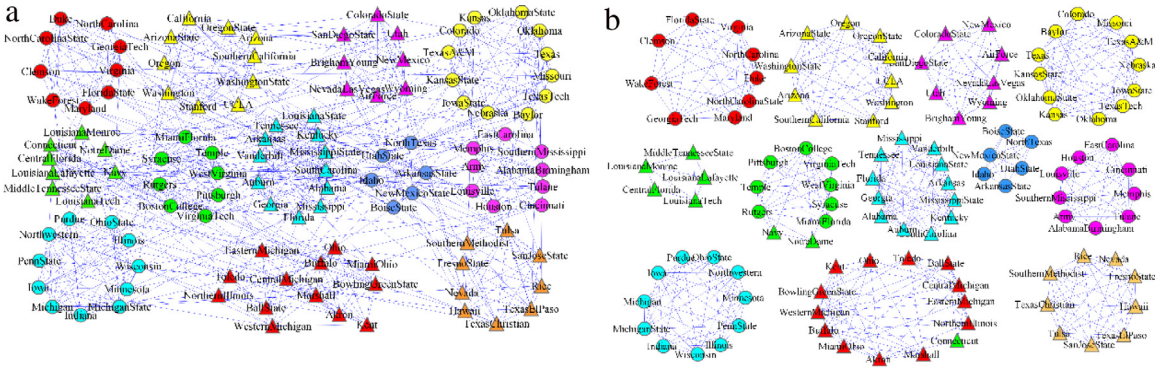
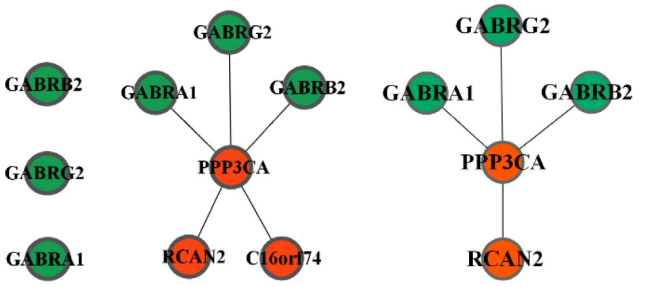


Fig. 7. Illustration of the real groups and the 2-club substructures obtained by DIVANC on football networks. Fig. 7(a) the football networks consisting of 12 groups; Fig. 7(b) the 2-club substructures obtained by DIVANC′ after removing 190 edges.



(a) Benchmark. (b) EPCA/EPD2/EPD2′. (c) DIVANC/DIVANC′.

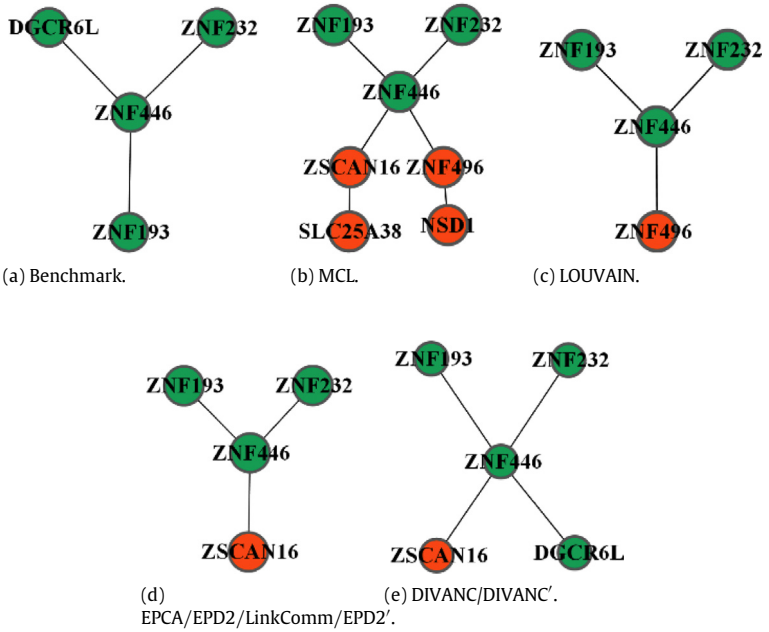
Fig. 8. Illustration of the benchmark and candidate complexes detected by the competing algorithms about GABAA receptor complex. Fig. 8(a) the benchmark of GABAA receptor complex; Fig. 8((b)–(c)) the corresponding candidate complexes detected by the non-overlapping algorithms EPCA, EPD2, DIVANC and overlapping algorithms EPD2′ and DIVANC′. The benchmark consisting of three isolated bright green proteins; among the detected candidate complexes the green proteins being the members of GABAA receptor complex and those in red color not. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We list the details of 4 sparse benchmarks and their corresponding candidate communities detected by the non-overlapping and overlapping algorithms in tables S2–S9 (Supplementary materials, Appendix C) and show them in Figs. 8–11. As described in tables S2 and S3, the density of GABAA receptor complex on *Hsa*HPRD is 0, thus it is really a challenge to detect its candidate complexes especially for the algorithms based on density. The algorithms based on density such as MCODE, COACH, ClusterOne and LinkComm even cannot obtain any valuable candidate ones which have common proteins with GABAA receptor complex. Among the algorithms, the neighborhood affinity scores (Appendix B) between the benchmark and the candidate complexes detected by DIVANC and DIVANC′ are the highest. We also list the details about DGCR6L-ZNF193-ZNF232-ZNF446-ZNF446 complex in tables S4, S5 and display them in Fig. 9, eEF-1 complex in tables S6, S7 and in Fig. 10, the 116th complex of the golden standard MIPS in tables S8, S9 and in Fig. 11. The best effectiveness of DIVANC and DIVANC′ in detecting sparse communities again verifies the value of developing algorithms based on the triad-rich substructures.

3.3. Testing practical performance of edge niche centrality and 2-hop overlapping strategy

3.3.1. Practical performance of edge niche centrality

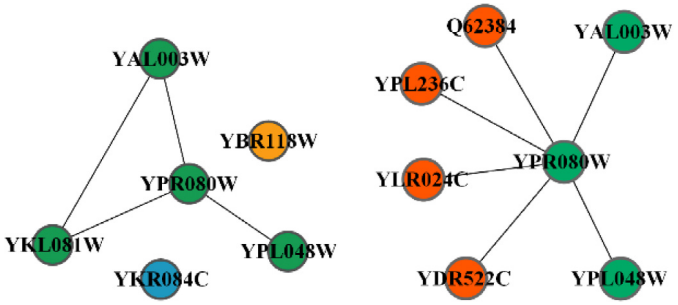
As we all know, edge centralities play an important role in edge division algorithms, thus in this section we want to compare edge niche centrality with edge P_4 centrality to show its advantages since the former is developed based on the



(a) Benchmark. (b) MCL. (c) LOUVAIN.

(d) EPCA/EPD2/LinkComm/EPD2′. (e) DIVANC/DIVANC′.

Fig. 9. Illustration of the benchmark and candidate complexes detected by the competing algorithms about DGCR6L-ZNF193-ZNF232-ZNF446-ZNF446 complex. Fig. 9(a) the benchmark of DGCR6L-ZNF193-ZNF232-ZNF446-ZNF446 complex; Fig. 9((b)–(e)) the corresponding candidate complexes detected by the non-overlapping algorithms MCL, LOUVAIN, EPCA, EPD2, DIVANC and overlapping algorithms LinkComm, EPD2′, DIVANC′. The benchmark consisting of 5 proteins coded by 4 genes, where the gene ZNF446 coded two proteins; among the detected candidate complexes the green color genes being the members of DGCR6L-ZNF193-ZNF232-ZNF446-ZNF446 complex and those in dark red color not. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

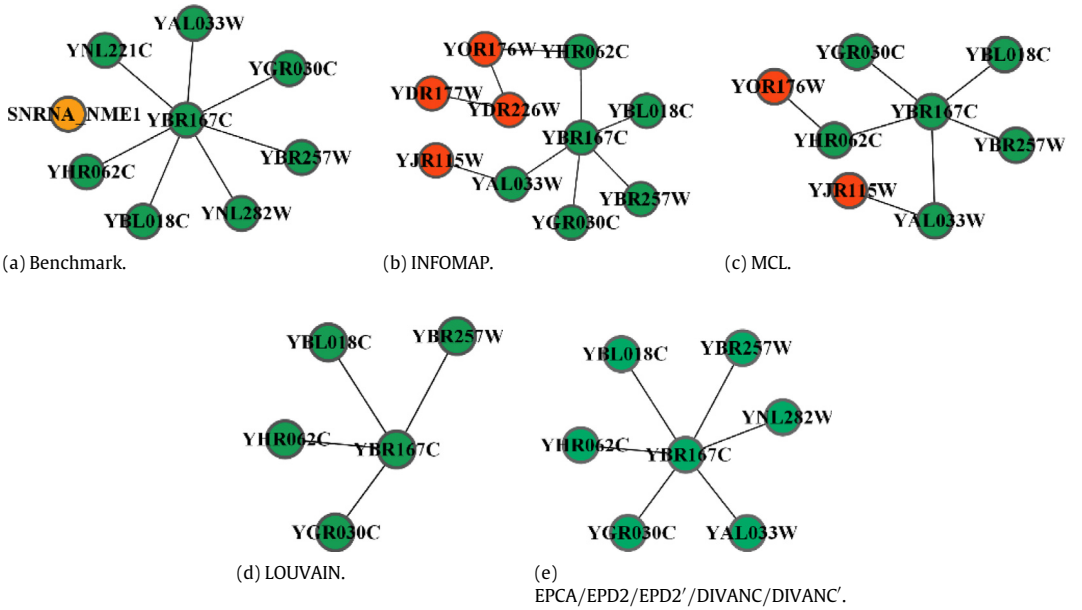


(a) Benchmark. (b) DIVANC/DIVANC′.

Fig. 10. Illustration of the benchmark and candidate complexes by the competing algorithms about eEF-1 complex. Fig. 10(a) the benchmark of eEF-1 complex; Fig. 10(b) the corresponding candidate complexes detected by DIVANC and DIVANC′. The benchmark consisting of 6 proteins, where the bright blue protein YKR084C is isolated proteins and YBR118W does not belong to the current input PPI networks since the incompleteness of datasets; among the detected candidate complexes the green proteins being the members of eEF-1 complex but those in dark red color not. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

latter. In fact, the comparisons between edge niche centrality and edge P_4 centrality are able to be implemented into the comparisons among their corresponding algorithms. We compare DIVANC with EPD2 (as introduced in above, EPD2 is a special edge division algorithm consisting of edge P_4 centrality and diameter 2 stop criterion, just scrabbled up only in order to compare the practical performances of edge niche centrality with edge P_4 centrality in detecting 2-club substructures).

As displayed in Tables 3 and 4, DIVANC detects 2151 candidate communities while EPD2 obtains 1942 ones on *Hsa*HPRD, and DIVANC detects 1128 candidate communities while EPD2 obtains 1015 ones on *Sce*DIP. Other than demonstrating their own relative indices of DIVANC and EPD2 in Tables 3 and 4, we also display the differences between them in this section. We see that a detected candidate community is able to match a golden standard complex or term if the score of neighborhood affinity (Appendix B) is equal or greater than 0.2 like in other parts of this paper. There are 249 candidate communities detected by DIVANC which cannot match any one of the 1942 ones obtained by EPD2. In other words, there are 249 candidate communities detected by DIVANC which cannot be obtained by EPD2 on *Hsa*HPRD, and likewise we can also detect 152 ones by DIVANC which cannot be obtained by EPD2 on *Sce*DIP. Surprisingly, among the 249 2-club substructures on *Hsa*HPRD, there are 71 ones which are able to match at least one community, and 16 of the 152 ones on *Sce*DIP are able to match at



(a) Benchmark. (b) INFOMAP. (c) MCL.

(d) LOUVAIN. (e) EPCA/EPD2/EPD2′/DIVANC/DIVANC′.

Fig. 11. Illustration of the benchmark and candidate complexes detected by the competing algorithms about MIPS (116th) complex. Fig. 11(a) the benchmark of MIPS (116th) complex; Fig. 11((b)–(e)) the corresponding candidate complexes detected by INFOMAP, MCL, LOUVAIN, EPCA, EPD2, DIVANC and EPD2′, DIVANC′. The benchmark consisting of 9 proteins, where the protein SNRNA_NME1 does not belong to the input PPI networks for the incompleteness of datasets; among the detected candidate complexes the green color proteins being the members of MIPS (116th) complex but those in dark red color not. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

least one community. Further, some of the candidate ones detected by DIVANC that cannot be detected by EPD2 are even able to match at least one community perfectly.

We display 8 candidate communities detected by DIVANC but cannot by EPD2 from *Hsa*HPRD and their corresponding matched benchmarks in Fig. 12 and more details in table S10; 4 those candidate communities from *Sce*DIP in Fig. 13 with more details in table S11. In addition to the official names of benchmarks and the gene members of benchmarks, we also list the scores of neighborhood affinity between the detected candidate communities and their own benchmarks. In the columns of 'Candidate communities' and 'Benchmark genes', we use bold fonts to label the common genes between candidate communities and benchmark genes. In Figs. 12 and 13, each benchmark consists of the genes in the area circled by dotted line and the candidate communities represented by the components consisting of the genes with red and green colors together. Among the genes in circled areas, the green genes are the common ones of candidate communities and benchmarks. While, the bright blue, yellow and purple genes are the ones which cannot be detected by DIVANC. Notably, the yellow genes are the ones which do not belong to the used PPI networks temporarily for the incompleteness of datasets and the bright blue ones are isolated proteins from the PPI networks, thus they will never be able to be detected by any algorithms. Only the purple ones are those missed by DIVANC. As we can see in Figs. 12 and 13, the genes of benchmarks cannot always be constructed as connected subnetworks also for the incompleteness property of the current PPI networks temporally. The fact that communities within the networks which are not always connected subnetworks and not to mention dense subnetworks, is just the challenges for community detection. In a word, those practical effectiveness comparisons between DIVANC and EPD2 reveal an obvious advantage of edge niche centrality over edge P_4 centrality.

3.3.2. Performance of 2-hop overlapping strategy

The algorithm DIVANC can be extended into overlapping version DIVANC′ by the proposed 2-hop overlapping strategy. In this section we mainly test the performance of 2-hop overlapping strategy in detail since it plays the role in detecting overlapping 2-club substructures. As described in Tables 3 and 4, whether on *Hsa*HPRD or on *Sce*DIP, DIVANC′ performs better than DIVANC overall. In other words, the better effectiveness of DIVANC′ justifies the value of the 2-hop overlapping strategy. The proposed 2-hop overlapping strategy not only produces new matched candidate communities, but also can improve their matching levels between detected candidate communities and their own benchmarks, and even makes some candidate ones to match benchmarks perfectly. Here we list 6 candidate communities detected by DIVANC which are further improved by the 2-hop overlapping strategy to match their own benchmarks perfectly in Fig. 14. The genes with green color are those detected by DIVANC, while the genes with red color are those detected additionally by the 2-hop overlapping strategy. Thus the overlapping algorithm DIVANC′ with 2-hop overlapping strategy can detect the candidate ones consisting of green genes and red genes together. As we list the neighborhood affinity scores in table S12, the candidate ones detected by DIVANC′ can match their own benchmarks perfectly. Although we demonstrate the significant performance of the 2-hop

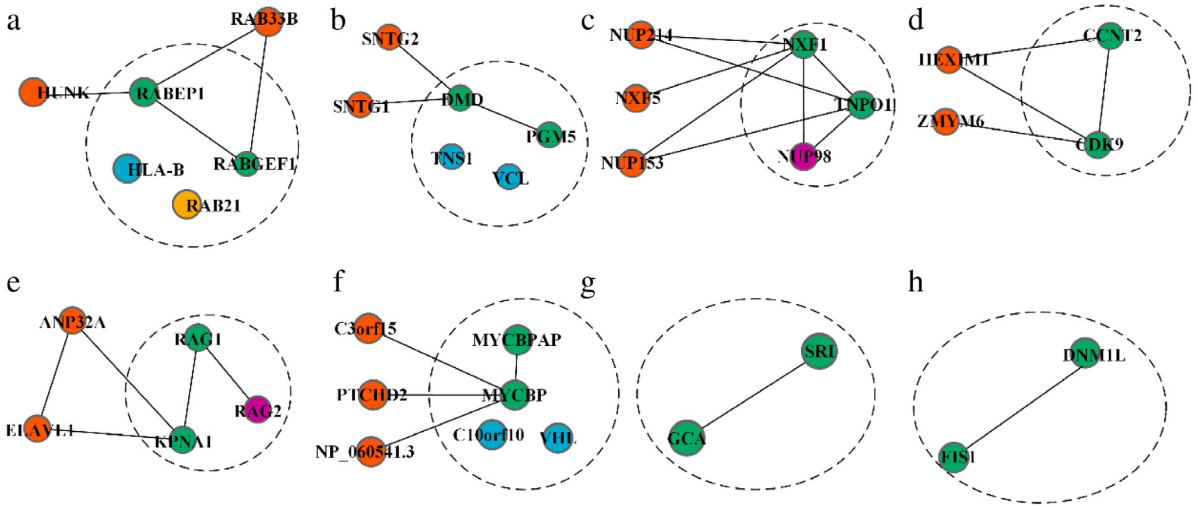


Fig. 12. Illustration of the candidate communities detected by DIVANC but cannot by EPD2 from *Hsa*HPRD. Fig. 12((a)–(h)) the 8 detected candidate communities (the connected subnetworks consisting of green and red proteins) and the benchmarks in the areas circled by dotted line as listed in table S10. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

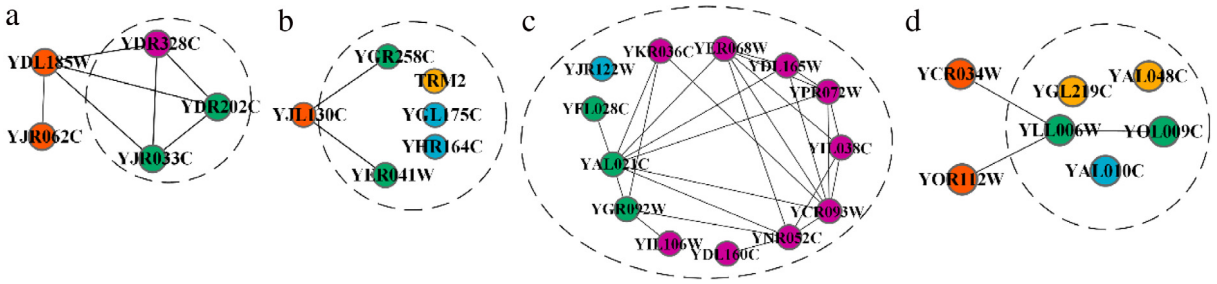


Fig. 13. Illustration of the candidate communities detected by DIVANC but cannot by EPD2 from *Sce*DIP. Fig. 13((a)–(d)) the 4 detected candidate communities (the connected subnetworks consisting of green and red proteins) and the benchmarks in the areas circled by dotted line as listed in table S11. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

overlapping strategy mainly by comparing the results of DIVANC and DIVANC′, the improved results of EPD2′ from EPD2 again verify the effects of 2-hop overlapping strategy as described in Tables 3 and 4 from another point of view. Thus the promotional effectiveness of EPD2′ over EPD2 also shows very well that the proposed 2-hop overlapping strategy has strong portability and can be widely used to turn other non-overlapping algorithms into overlapping ones.

4. Conclusions and discussion

In this work, we aim to overcome the challenge that traditional definitions cannot characterize intrinsic features of communities comprehensively. We develop a new framework by incorporating the novel assumption of triad-rich substructures, defining 2-club substructures, designing the effective algorithm DIVANC to detect non-overlapping and overlapping candidate communities that have desired graph-theoretic properties. To verify the effectiveness of triad-rich substructures, we compare DIVANC with existing algorithms on PPI networks, LFR synthetic networks and football networks. The experimental results reveal DIVANC outperforms most other exiting algorithms significantly and, in particular, can detect sparse communities.

In a future study, we will attempt to study the possible applications of 2-club subclasses on complex networks from the viewpoint of graph theory since 2-club substructures have interesting internal structures.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61532014, 91530113, 61432010, 61303122, 61303118, 61402349, 71401130; and the Fundamental Research Funds for the Central Universities under Grant No. BDZ021404.

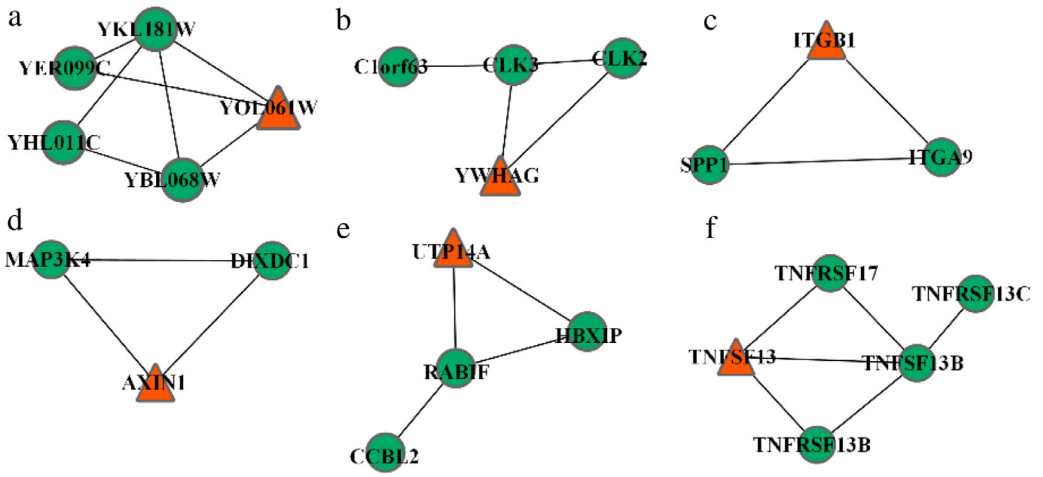


Fig. 14. Illustration of the communities detected by DIVANC′ matching their own benchmarks perfectly. Fig. 14((a)–(f)) the 6 communities as listed in table S12, where the red triangle proteins being those searched by the 2-hop overlapping strategy and together with the non-overlapping green circle proteins matching their own benchmarks perfectly. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Appendix A. The details about two protein complexes

Nuclear origin of replication recognition complex

A multisubunit complex that is located at the replication origins of a chromosome in the nucleus [20].

GID complex

A protein complex with ubiquitin ligase activity that is involved in proteasomal degradation of fructose-1,6-bisphosphatase (FBPase) and phosphoenolpyruvate carboxykinase during the transition from gluconeogenic to glycolytic growth conditions [21].

Appendix B. The details about relative indices

The numbers of matching communities

The numbers of detected candidate communities matching at least one community are important indices for comparison. A detected candidate community M_A with $|V_{M_A}|$ proteins or genes is thought to match with a community M_B with $|V_{M_B}|$ proteins or genes if the score of neighborhood affinity

$$NA(M_A, M_B) = \left|V_{M_A} \bigcap V_{M_B}\right|^2 / \left(|V_{M_A}| \times |V_{M_B}|\right) \geq \omega, \tag{B.1}$$

where the threshold ω is usually set as 0.2 or 0.25 in Refs. [43,46].

Accuracy score

We mainly make use of accuracy score (Acc) to evaluate the performance of various algorithms on protein complex detection. It is the geometric mean of cluster-wise sensitivity (Sn) and cluster-wise positive predictive value (PPV) [46]. Given r detected and s reference complexes, let t_{ij} represent the number of proteins that exist in both detected complex i and reference complex j, and w_j represents the number of proteins in reference complex j. Then Sn and PPV are defined as $Sn = \sum_{j=1}^{s} \max_{i=1,\ldots,r}\{t_{ij}\} / \sum_{j=1}^{s} w_j$, $PPV = \sum_{i=1}^{r} \max_{j=1,\ldots,s}\{t_{ij}\} / \sum_{i=1}^{r}\sum_{j=1}^{s} t_{ij}$ respectively. Since Sn can reach its maximum by grouping all proteins in one complex, whereas PPV can be maximized by putting each protein in its own complex, we use their geometric mean

$$Acc = \sqrt{Sn \times PPV}, \tag{B.2}$$

as 'accuracy' to balance these two indices [43,46], where the higher Acc scores mean the better results.

F-measure

To investigate the performance of competing algorithms in detecting GO terms, we can compute the indices of F-measure [50]. If the neighborhood affinity score between a detected candidate GO term p and a real GO term rg, $NA(p, rg) \geq \omega$, they are considered to be matched with each other. Assuming PC as the set of candidate ones detected by an algorithm, RG

as the set of real GO terms, N_{cp} indicating the number of candidate ones which can match at least one real GO term and N_{crg} representing the number of real GO terms that match at least one candidate term, we obtain precision (P) and recall (R) as follows [50]: $N_{cp} = |\{p\,|p \in PC, \exists\ rg \in RG, NA(p, rg) \geq \omega\}|$, $N_{crg} = |\{rg\,|rg \in RG, \exists p \in PC, NA(p, rg) \geq \omega\}|$, $P = N_{cp}/\,|PC|$, $R = N_{crg}/\,|RG|$. F-measure (F) as the harmonic mean of precision and recall, thus we have

$$F = (2 \times P \times R)\,/\,(P + R)\,. \tag{B.3}$$

Percentage of matched GO terms

Percentage of matched GO terms which are considered to be the percentage of the GO terms which are correctly matched to at least one of the identified candidate GO terms [34,35].

Maximum matching ratio

Here we also use a measure called maximum matching ratio (MMR) [5] to evaluate relative algorithms on detection of protein complexes and GO terms. The MMR builds on maximal matching in a bipartite network, in which the two sets of vertices represent the reference and detected community, respectively, and an edge connecting a reference community with a detected one is weighted by the score of neighborhood affinity introduced in (Eq. (B.1)). We select the maximum weighted bipartite matching on this network; that is, we chose a subset of edges such that each of detected and reference communities is incident on at most one selected edge and the sum of the weights of such edges is maximal. The chosen edges then represent an optimal assignment between reference and detected communities such that no reference community is assigned to more than one detected community and vice versa. The MMR between the detected and the reference community set is then given by the total weight of the selected edges, divided by the number of reference communities. MMR offers a natural, intuitive way to compare detected communities with a gold standard and it explicitly penalizes cases when a reference community is split into two or more parts in the predicted set, as only one of its parts is allowed to match the correct reference community.

Normalized mutual information

Normalized mutual information (NMI) is well known for evaluating community detection algorithms. In this paper we use the version of NMI_{MGH} [48] to assess the similarities between detected results and golden standards on football networks and the series of LFR synthetic networks. Its definition is demonstrated as

$$NMI_{MGH} = I(X : Y)/\max(H(X), H(Y)), \tag{B.4}$$

where $I(X : Y)$ is the mutual information, $H(X)$, $(H(Y))$ the unconditional entropy of cover X, (Y). More details can be found in original Refs. [48,49].

Appendix C. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.physa.2016.10.021.

References

[1] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (2010) 75–174.
[2] A.J. Enright, S. Van Dongen, C.A. Ouzounis, An efficient algorithm for large-scale detection of protein families, Nucleic Acids Res. 30 (2002) 1575–1584.
[3] M. Rosvall, C.T. Bergstrom, Maps of random walks on complex networks reveal community structure, Proc. Natl. Acad. Sci. 105 (2008) 1118–1123.
[4] G.D. Bader, C.W.V. Hogue, An automated method for finding molecular complexes in large protein interaction networks, BMC Bioinformatics 4 (2003) 2.
[5] T. Nepusz, H. Yu, A. Paccanaro, Detecting overlapping protein complexes in protein-protein interaction networks, Nat. Methods 9 (2012) 471–472.
[6] Y.-Y. Ahn, J.P. Bagrow, S. Lehmann, Link communities reveal multiscale complexity in networks, Nature 466 (2010) 761–764.
[7] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. Theory Exp. 2008 (2008) P10008.
[8] A. Lancichinetti, F. Radicchi, J.J. Ramasco, S. Fortunato, Finding statistically significant communities in networks, PLoS One 6 (2011) e18961.
[9] J. Zhang, K. Zhang, X. Xu, K.T. Chi, M. Small, Seeding the kernels in graphs: Toward multi-resolution community analysis, New J. Phys. 11 (2009) 113003.
[10] D. Deritei, Z.I. Lázár, I. Papp, F. Járai-Szabó, R. Sumi, L. Varga, E.R. Regan, M. Ercsey-Ravasz, Community detection by graph Voronoi diagrams, New J. Phys. 16 (2014) 063007.
[11] S. Jia, L. Gao, Y. Gao, J. Nastos, Y. Wang, X. Zhang, H. Wang, Defining and identifying cograph communities in complex networks, New J. Phys. 17 (2015) 013044.
[12] F. Radicchi, A paradox in community detection, Europhys. Lett. EPL 106 (2014) 38001.
[13] J. Yang, J. Leskovec, Overlapping community detection at scale: a nonnegative matrix factorization approach, in: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, ACM, 2013, pp. 587–596.
[14] J. Yang, J. Leskovec, Defining and evaluating network communities based on ground-truth, Knowl. Inf. Syst. 42 (2013) 181–213.
[15] D. Hric, R.K. Darst, S. Fortunato, Community detection in networks: Structural communities versus ground truth, Phys. Rev. E 90 (2014) 062805.
[16] J. Yang, J. Leskovec, Structure and overlaps of ground-truth communities in networks, ACM Trans. Intell. Syst. Technol. TIST 5 (2014) 26.
[17] R. Balasubramanyan, W.W. Cohen, Block-LDA: Jointly modeling entity-annotated text and entity-entity links, in: SDM Vol. 11, SIAM, 2011, pp. 450–461.
[18] Y. Ruan, D. Fuhry, S. Parthasarathy, Efficient community detection in large networks using content and links in: Proceedings of the 22nd International Conference on World Wide Web (International World Wide Web Conferences Steering Committee), 2013, pp. 1089–1098.
[19] S. Pool, F. Bonchi, M. van Leeuwen, Description-driven community detection, ACM Trans. Intell. Syst. Technol. TIST 5 (2014) 28.
[20] M. Balasov, R.P. Huijbregts, I. Chesnokov, Functional analysis of an Orc6 mutant in Drosophila, Proc. Natl. Acad. Sci. 106 (2009) 10672–10677.

[21] O. Santt, T. Pfirrmann, B. Braun, J. Juretschke, P. Kimmig, H. Scheel, K. Hofmann, M. Thumm, D.H. Wolf, The yeast GID complex, a novel ubiquitin ligase (E3) involved in the regulation of carbohydrate metabolism, Mol. Biol. Cell 19 (2008) 3323–3333.
[22] D.M. Wolf, A.P. Arkin, Motifs, modules and games in bacteria, Curr. Opin. Microbiol. 6 (2003) 125–134.
[23] E. Yeger-Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R.Y. Pinter, U. Alon, H. Margalit, Network motifs in integrated cellular networks of transcription–regulation and protein–protein interaction, Proc. Natl. Acad. Sci. USA 101 (2004) 5934–5939.
[24] I. Albert, R. Albert, Conserved network motifs allow protein–protein interaction prediction, Bioinformatics 20 (2004) 3346–3352.
[25] S.-H. Yook, Z.N. Oltvai, A.-L. Barabási, Functional and topological characterization of protein interaction networks, Proteomics 4 (2004) 928–942.
[26] L. Salwinski, C.S. Miller, A.J. Smith, F.K. Pettit, J.U. Bowie, D. Eisenberg, The database of interacting proteins: 2004 update, Nucleic Acids Res. 32 (2004) D449–D451.
[27] T.K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, Human protein reference database—2009 update, Nucleic Acids Res. 37 (2009) D767–D772.
[28] A. Lancichinetti, S. Fortunato, F. Radicchi, Benchmark graphs for testing community detection algorithms, Phys. Rev. E 78 (2008) 046110.
[29] A. Lancichinetti, S. Fortunato, Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities, Phys. Rev. E 80 (2009) 016118.
[30] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. 99 (2002) 7821–7826.
[31] T.S. Evans, Clique graphs and overlapping communities, J. Stat. Mech. Theory Exp. 2010 (2010) P12037.
[32] H.-W. Mewes, C. Amid, R. Arnold, D. Frishman, U. Güldener, G. Mannhaupt, M. Münsterkötter, P. Pagel, N. Strack, V. Stümpflen, MIPS: analysis and annotation of proteins from whole genomes, Nucleic Acids Res. 32 (2004) D41–D44.
[33] E.L. Hong, R. Balakrishnan, Q. Dong, K.R. Christie, J. Park, G. Binkley, M.C. Costanzo, S.S. Dwight, S.R. Engel, D.G. Fisk, Gene Ontology annotations at SGD: new data sources and annotation methods, Nucleic Acids Res. 36 (2008) D577–D581.
[34] Y.-K. Shih, S. Parthasarathy, Identifying functional modules in interaction networks through overlapping Markov clustering, Bioinformatics 28 (2012) i473–i479.
[35] Y. Wang, X. Qian, Functional module identification in protein interaction networks by interaction patterns, Bioinformatics 30 (2014) 81–93.
[36] S. Kikugawa, K. Nishikata, K. Murakami, Y. Sato, M. Suzuki, M. Altaf-Ul-Amin, S. Kanaya, T. Imanishi, PCDq: human protein complex database with quality index which summarizes different levels of evidences of protein complexes predicted from H-Invitational protein-protein interactions integrative dataset, BMC Syst. Biol. 6 (2012) S7.
[37] A. Ruepp, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, M. Stransky, B. Waegele, T. Schmidt, O.N. Doudieu, V. Stümpflen, H.W. Mewes, CORUM: the comprehensive resource of mammalian protein complexes, Nucleic Acids Res. 36 (2008) D646–D650.
[38] B. Serrour, A. Arenas, S. Gómez, Detecting communities of triangles in complex networks using spectral optimization, Comput. Commun. 34 (2011) 629–634.
[39] D.G. Corneil, H. Lerchs, L.S. Burlingham, Complement reducible graphs, Discrete Appl. Math. 3 (1981) 163–174.
[40] S. Jia, L. Gao, Y. Gao, H. Wang, Anti-triangle centrality-based community detection in complex networks, IET Syst. Biol. 8 (2014) 116–125.
[41] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, Defining and identifying communities in networks, Proc. Natl. Acad. Sci. USA 101 (2004) 2658–2663.
[42] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (2005) 814–818.
[43] M. Wu, X. Li, C.-K. Kwoh, S.-K. Ng, A core-attachment based method to detect protein complexes in PPI networks, BMC Bioinformatics 10 (2009) 169.
[44] G. Csardi, T. Nepusz, The igraph software package for complex network research, Int. J. Complex Syst. (2006) 1965.
[45] A.T. Kalinka, P. Tomancak, Linkcomm: an R package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type, Bioinformatics 27 (2011).
[46] X. Li, M. Wu, C.-K. Kwoh, S.-K. Ng, Computational approaches for detecting protein complexes from protein interaction networks: a survey, BMC Genomics 11 (2010) S3.
[47] C. Yamasaki, K. Murakami, J. Takeda, Y. Sato, A. Noda, R. Sakate, T. Habara, H. Nakaoka, F. Todokoro, A. Matsuya, T. Imanishi, T. Gojobori, H-InvDB in 2009: extended database and data mining resources for human genes and transcripts, Nucleic Acids Res. 38 (2010) D626–D632.
[48] A.F. McDaid, D. Greene, N. Hurley, Normalized mutual information to evaluate overlapping community finding algorithms 2011. ArXiv Prepr. arXiv11102515.
[49] A. Lancichinetti, S. Fortunato, J. Kertész, Detecting the overlapping and hierarchical community structure in complex networks, New J. Phys. 11 (2009) 033015.
[50] Y.-R. Cho, W. Hwang, M. Ramanathan, A. Zhang, Semantic integration to identify overlapping functional modules in protein interaction networks, BMC Bioinformatics 8 (2007) 265.